Supervised : 정답 있는

Unsupervised : 정답 없는

CNN (convolutional neural network) : image classification 다양한 이미지로 학습하여 다음 샘플에 대해 예측하게 함, 이미지 구분

RNN (recurrent neural network) (Long Short Term Memory) : natural language processing, text, sequence 를 번역, 순서

GAN (generative adversarial network) : 사진, 오디오 생성

Classification – yes or no 맞추기 기본은 binary

Estimation – 정확한 값

Classifiers 만드는 순서

데이터 로드

데이터를 트레이닝, 테스트로 나눔

모델 import (scaler 등 가져오기)

모델 트레이닝

테스트 데이터로 테스트 (예측)

모델 evaluate : 예측을 제대로 했는지 확인

Decision tree 에서는 기준점 (%) 를 미리 시작하기 전에 정해놓음

여러가지 feature에서 뭐가 root가 될지는 아무도 모름 다 해봐야 함

Decision tree 만들기 (induction) 🡪 model 제작 🡪 적용 (deduction)

Key of induction

* feature들이 많으면 순서를 어떻게 정할 것 인가
* split 기준을 어떻게 세울 것인가
* 언제 tree를 멈추고 결정할 것인가

Split 할 때 multi way, binary 로 할 수 있다

Ordinary 에서는 순서를 건너뛰어서 묶지 않는다

Numerical data에서는 bins를 설정하여 나누거나 (범위 지정), 기준(계산을 통해서 산출됨)을 정해서 위, 아래로 나눌 수 있다.

뭐가 가장 좋음? Determining the best split

Purity (homogeneous)를 이용한다

Degree of impurity가 낮을수록 pure 함

Accuracy 50%? 쓰레기 failure 90%는 나와야 한다.

Decision tree를 멈추기 위한 요인 – impurity

Impurity measure

Entropy, gini index, classification error, variance reduction

Information gain의 경우 branch가 많은 경우 우선권을 받게 되어서 문제가 생긴다

Decision stump : root만 있는 것

k-nearest : uclid 거리, 사실상 training이 없음, 전체 dataset 대상으로 거리를 모두 구함

k = 홀수 걍 당연한 것 (가까운 것에 가중치를 준다면 tie도 해결 가능하지만 대체로 홀수)

아까 decision tree + k-nearest 까지 하기

제각기 coding 하고 discuss